An Approach for Mining Accumulated Crop Cultivation Problems and their Solutions

Samhaa R. El-Beltagy*, Ahmed Rafea[γ], Said Mabrouk[Ω] and Mahmoud Rafea[Ω]

*Cairo University, [γ]The American University in Cairo, [Ω]The Central Lab for Agricultural Expert Systems
*Faculty of Computers and Information, 5 Dr. Ahmed Zewail Street, 12613, Orman, Giza, Egypt
[γ]Computer Science Department, AUC Avenue, P.O. Box 74, New Cairo 11835, Egypt.
[Ω]Ministry of Agriculture and Land Reclamation, Giza, Egypt.
samhaa@computer.org, rafea@aucegypt.edu, said51@windowslive.com, and mahmoud@claes.sci.eg

Abstract

This paper presents an approach for mining agricultural problems that have been accumulated in a textual database over a period of 5 years. The problems, which are accompanied by their solutions, offer a wealth of knowledge that can be used by decision makers, researchers, and farmers alike. However, this wealth of knowledge can not be unlocked without a) representing these problems in a structured format, and b) applying algorithms that can summarize and analyze this information. Towards the achievement of the first goal, a multi-faceted object extraction methodology is presented, and for the achievement of the second, association rules are employed. As a proof of concept, the tool was applied of a set of weed problems. The presented methodology can be modified to work with any help and support textual database where both problems and their solutions are present.

Introduction

As a developing country with limited expertise, Egypt has realized early on the importance of employing AI and information technologies for promoting Agricultural productivity. One of the by-products of this realization was the establishment of the Central Laboratory for Agricultural Expert Systems (CLAES 2008) another was the development of the Virtual extension and Research Network (VERCON 2006). In the Agricultural sector, fast access to problem solutions can greatly influence productivity. One of VERCON's subsystems is the "Farmers Problems Database" which was created in order to address farmer problems that cannot be anticipated in advance. The idea of this component was a simple one: through a web interface, a user can input his/her problem using some basic meta-data descriptors that specify his/her location, the nature of the problem being entered, and the crop for which this problem has been encountered. S/he would then go on to enter a free text description of the problem listing all concerns and additional important parameters within the text itself. The problem would then be forwarded to a researcher who would reply in free text telling the farmer what s/he should do to address the problem. The idea was that someone with a problem can search this resource for a similar problem to his/hers and readily find a response to his/her concern. If such a solution could not be found, then the person can then post their problem, have it answered by an expert, and avail the complaint and its solution to other users of the system. Over a period of five years, this valuable resource has grown to contain 10,000+ problems and their solutions all stored in a textual database. With this growth, redundant problem entry started to take place as users started finding it more difficult to locate problems similar to theirs. With this growth also, the opportunity and potential of mining and extracting information from this resource was identified with the following objectives in mind:

- First, patterns and relations can be discovered and used to enhance the utilization of this valuable resource. The discovered patterns and relations may point to certain types of widespread problems and pressing needs of people living in rural areas. Consequently, decision makers could be able to take necessary actions to tackle these pressing problems and needs of poor communities, and direct development plans toward these needs
- Second, solutions given for similar problems, by different experts or by the same expert at a different time can be analyzed in terms of their similarities and differences. Inconsistencies can then be resolved

through statistical consensus, which can then be validated by a domain expert.

- Third, patterns of problems and their solutions can be created and used to classify new problems and provide solutions without the need for domain experts.
- Fourth, outdated recommendations that contain prohibited material can be easily identified and removed from the database.
- Fifth, users using the complaint database can easily locate problems that are similar to theirs just by entering a free text description of their complaint.

But in order to achieve any of these objectives, both the complaints and the solutions need to be represented in a structured format. Since the problems and their solution are presented mainly in text, converting that text into a more structured format through the use of text mining techniques has to be carried out first. The challenge lies in the identification of the complaint object, the features describing this object, and the complaint features that specify the focus of a given complaint, and consequently the discovery of similar complaints written in different styles.

This paper presents the initial approach and results of carrying out this task on a subset of problems extracted from the VERCON problem database. The approach can be modified to work with any help and support repository.

The rest of this paper is organized as follows: section 2 presents a brief review of related work. Section 3, provides a closer look at the problem being addressed. Section 4 provides an overview of the implemented object extraction process. Section 5, briefly outlines an implemented prototype that makes use of the extraction process as well as of association rules. Finally, section 6 concludes this paper and presents future work.

Related Work

This work aims at analyzing growers' complaints to come up with causes and features of these complaints and their solutions. This area of research has some similarity with opinion mining which appeared recently to assist customers in product reviews before their purchase as it became very difficult to go through the huge amount of reviews available on the web. The growers' complaints represent negative opinions about agricultural objects such as soil, water, climate, plant and others. We can make an analogy between these multiple objects hidden in the complaints and opinion mining of products reviews.

Most opinion mining systems rely on identifying the product features and their possible associated opinions. For example, the feature "display" is associated with opinion words like "bright", "dark", and "clear" for a mobile phone product (Shi and Chang 2006). (Hu and Liu 2004) proposed an automatic way to extract product features from English product reviews using association rule mining (Agrawal and Srikant 1994). (Berland and Charniak 1999) proposed a method to extract "part-of" type features using possessive constructions and prepositional phrases from news corpora. (Yi et al. 2003) extracted both "part-of" and "attribute-of" type features from online reviews. (Popescu and Etzioni 2005), extracted explicit features for a given product class from parsed review data, and used PMI assessment to evaluate each candidate feature. (Liu, Wu, and Yao 2006) presented a method based on identifying all the domain-related phrases and then divided them into features and products.

Extracting opinion words that describe a certain feature is a challenging problem as is the identification of product features. The opinion word dictionary is typically generated by collecting adjectives expressing positive or negative opinions and then automatically searching for synonyms and antonyms using a semantic lexicon such as WordNet (Miller el al, 1990) (Hu and Liu 2004). (Liu, Hu, and Cheng 2005) proposed an Opinion Observer, from which users can clearly see the advantage and weakness of each product in the minds of other consumers. Their work has boosted the development of new techniques and systems for opinion analysis.

However, this work has to address the handling of multiple objects and not a single product which is the focus of most of the current research done in opinion mining. Using an ontology to discover these multiple objects, identifying features of multiple objects, and extracting the words that specify the complaints, outline the main contributions of this work.

Problem Analysis

By examining a sample of complaints and solutions entered into the VERCON problem database, it was found that a single complaint may contain 1 or more primary complaint objects (explicitly or implicitly specified) often supported with additional complaint features. An implicitly implied complaint object, is one that is not explicitly entered by the complainer, and which is usually deduced and stated by the expert answering the user's query. For example, a user entering a complaint about a disease may not even know what the disease is and may simply describe it in terms of symptoms occurring on various plant parts. An expert would take this, and provide the disease name as well as advice as to how to address this problem. The disease in this case, is the primary object of the complaint, and the symptoms entered by the user are the supporting features. Each of these features typically describe a specific plant part, and the extraction process has to differentiate between these. An expert often describes alternative solutions rather than one. So extracting each of these

without confusing features associated with each, is another one of the extraction challenges. A close look at various entered problems and solutions revealed that farmers and experts alike, enter problems and solutions with varying degrees of details.

The following is an example of typical complaint retrieved from the textual database of VERCON. The complaint has been translated from Arabic to English:

There are spots on the leaves and on the spikes which have a cotton like texture and which turn to grey in some areas within the planted 25 feddan land.

The main object of the complaint (the disease) is not apparent in this text, but we need to identify that the descriptor complaint objects are the leaf and spike, and that the complaint feature describing either is the spots that have cotton like texture and a color that is changing to gray. More formally, the attributes of the spot feature will be as follows: color = grey, and texture = cotton like. So if we find the same features in another complaint, we can predict that these two complaints are similar. However, in real life it is rare to have two problems with the exact same features. For example, the following is a complaint that is similar to the one given above but which has different wordings and features:

There are white, non-uniform spots with cotton like texture on the lower surface of plant leaves.

Here the descriptor complaint object is the leaf, and the feature associated with this descriptor is spots which have the following attributes: (color = white, texture = cotton like, location = lower surface, and distribution= non-uniform). The two problems partially match in terms of problem features as they both have spots on leaves characterized by the a cotton like texture. However, this information on its own is not enough to deduce that the two problems are actually similar. Looking at the solutions for both problems, we find that the solution is the same. We also find the experts responding to the query both name "Powdery mildew" (which we consider as the main object of the complaint) as the main cause of the complaint. So, the solution can actually help in the production of generalization rules that can both determine problems that are similar to each other, as well as aid users with future complaints, in finding solutions to their problems by either storing a "standard solution", or simply displaying similar problems and their offered solutions.

Overview of the Extraction Process

In the textual complaints database, metadata is provided to classify problems according to the issues they deal with. For example, typical classifications include "weeds", "diseases", "pests", "fertilization", and "irrigation". For each of these categories, a hand crafted extraction template is created. A filled in template would represent a problem and its solution. So basically, a template specifies objects that can be extracted from these and their types. Objects to extract are defined as one of five possible types: Named entities, time based entities, numbers, percentages and rates. In this work, the use of an ontology is instrumental. An ontology plays a major role in the identification of agricultural objects which are the named entities of concern in this work. The ontology not only includes specific objects, but also their associated features and the possible attributes for these. For the extraction task, we have used a similar methodology to the one we had employed for the purpose of text segments annotation which we also carried out using an ontology (El-Beltagy, Hazman, and Rafea 2007). Initially, all ontology entries are read, stemmed, and stored in their stemmed form. Stemming of Arabic is far more complicated than that of English because of its inflected nature. Since accuracy of this step can affect the overall accuracy of the system, we developed our own stemmer for carrying out this task (El-Beltagy and Rafea 2009). When parsing an input complaint (which is actually the complaint and its solution), the complaint's text is scanned word by word. A search for each stemmed version of word is then carried out within stemmed version of the ontology. If a match is found, then the concept/object associated with the word, is said to have been identified. For example, when the word "دنيبه" is encountered and searched within ontology, it will be found to be associated with the 'Weed' object. The object and the location where it was encountered (word number and sentence number) are both stored. This is important as we define a context window for the identification of related features. For example, a pesticide has three related features which are: concentration (percentage), time_of_Application(time based entity), and rate_of_application(rate). So, if a pattern matching any of these features is detected in the object's vicinity, then its associated feature can be extracted and the value of its attribute set.

A number is easy to extract as it usually follows a known number format. Time based entities can appear as a number followed by a unit of time (for example, 1 day, 2 months, 3 years), or they can simply be represented in pure text (for example "after a week"). A percentage entity is usually a number followed by a percentage mark and is often a property of a named entity. A rate refers to an amount of one thing considered in relation to a unit of another thing (Kilos per Feddan, dollars/hour, etc). Regular expressions are employed for the extraction of these four entity types.

When defining an extraction template, a user can indicate that certain objects are context objects, and can define context windows for the extraction of their associated features. Other objects which are features of this object are associated with this object through a context field which lists this object as their context. A

user can also specify a certain object as the main object of a complaint (MOC) and another as a main object of a solution (MOS). During the extraction process, a tuple for each complaint and its solution is created. This tuple, is replicated each time a new MOC or MOS is encountered. Figure1 , shows a simplified representation of the extraction process.

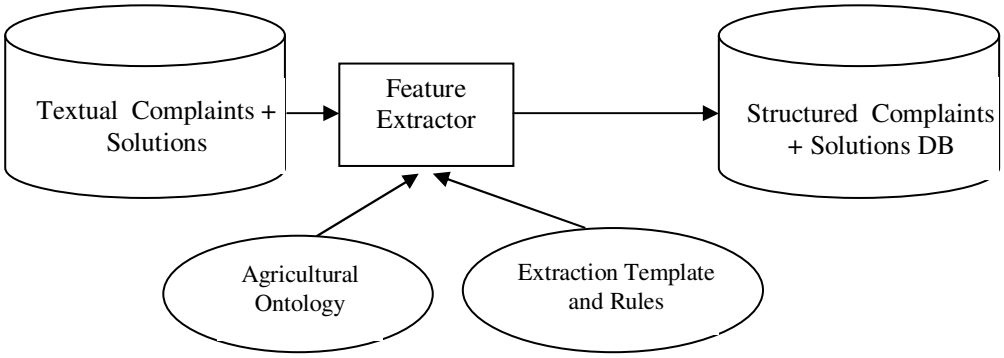

Figure1 : An overview of the extraction process

After the problems and their solutions are stored in a structured database, further analysis can be carried out on them using for example association rules.

Experimental Testbed

A subset of problems dealing with weed associated problems was selected for the purpose of experimentation with the proposed approach. The reason this specific class of problems was selected is due to the fact that they represent the biggest class of problems in the textual database. Association rules where applied on the database created as a result of the extraction process. Various combinations of features were generated using up to 4 item sets. The minimum support value, was set to 10% in case if one itemset, and 5% for all other itemsets. Examination of the results, led to the selection of a set of potentially useful patterns and meaningful relationships. These are outlined as follows:

1. **The most frequently occurring weeds and their occurrence frequency.**
2. **The distribution of weeds over governorates.**
3. **The distribution pattern of weed problems among planting methods**.
4. **The most commonly used herbicides and their occurrence frequency.**
5. **Relationship between a certain weed and a specific herbicide**.
6. **Relationship between the control method and control time.**
7. **Relationship between herbicides and control times.**
8. **Breakdown of weeds into (wide and narrow weeds) and their occurrence frequency**
9. **The relationship between generalized weeds and herbicides.**

Figure 2 shows a sample report representing the relationship between a given weed and a given herbicide.

We were also able to obtain a list of weeds that are always covered by recommendations issued by the Ministry of Agriculture as well as another of prohibited herbicides. When producing any of the above outputs, weeds that have been reported by farmers, but which are not covered by recommendations are highlighted in red. These probably refer to problems that decision makers aren't even aware of. Similarly, whenever a herbicide recommended by an expert matches with one that has been prohibited, the herbicides name is highlighted in red. But in this case also, all recommendations advising the use of this herbicide are relocated from VERCON's database, such that users may no longer see this herbicide being advised.

Relationship between a certain weed and herbicide (Threshold 5%)

Weed	Herbicide	Concentration	Rate	Support %	Confidence %
دنيبة	ساتيرن	50%	2 لتر/فدان	13	44
عجيرة	ساتيرن		2 لتر/فدان	9.2	38
دنيبة	كفروساتيرن		2 لتر/فدان	6.5	22
البوركية	ساتيرن	50%		6	48
سعد	بازجران		1.5 لتر/فدان	5.4	42

Back

Figure 2: A sample report representing the relationship between a given weed and a given herbicide

This result, satisfies the first objective of our work as stated in the introduction. Work is underway to satisfy the other 4 objectives.

Conclusion and Future Work

This work has briefly overviewed our used methodology for transforming the free text of Agricultural complaints and their solutions, to a structured format. Representing textual complaints and their solutions in a structured format and analyzing this information can be useful to growers, researchers and decision makers alike. Discovered patterns and relations, can guide researchers to new previously unknown knowledge that they can further investigate. They can also point decision maker to problems that they are unaware of and tell them how serious these problems are. Being able to spot inconsistence in expert advice and resolve this, can lead to less confusion among users of the "Farmers Problem Database" and enhance its credibility. More importantly, it will guide users to actually apply appropriate solutions to their problems, which will directly affect their productivity. Having a means for converting free unstructured text, into a structured form, means that growers can simply enter their problems in free text and as a result obtain solutions to problems that are similar to theirs. Currently, users of the "farmers problems database", are no longer capable to carrying out appropriate search on the database because of its size, so they simply enter their problem and wait for it to be answered by an expert when the answer is often already in the database. So having such a search interface will accelerate the response time to a problem, which can again directly affect productivity.

This work has also shown how the use of the association rules on extracted information can result in the production of useful patterns and relations. More work is currently underway in order to reach the full potential of this extraction process as outlined by our objectives. Work is also currently underway to avail extracted patterns and relations through a web based system, as well as on expanding our initial prototype to work with other problem categories.

Acknowledgments

This work has been supported by the Egyptian Science and Technology for Development Fund.

References

Agrawal R., and Srikant, R. 1994. Fast Algorithm for Mining Association Rules. In *Proceedings of* VLDB'94,. 487-499. Santiago, Chile.

Berland, M. and Charniak, E. 1999. Finding parts in very large corpora. In *Proceedings of the 37th ACL Conference,* 57–64. College Park, Maryland: Association for Computational Linguistics.

CLAES, 2003. http://www.claes.sci.eg/.

El-Beltagy, S., Hazman, M., and Rafea, A. 2007. Ontology Based Annotation of Web Document Segments. In *Proceedings of the 22nd Annual ACM Symposium on Applied Computing (SAC'07),* 1362-1367, Seoul, Korea.

El-Beltagy, S. , Hazman, M., and Rafea, A. 2009. A Framework for the Rapid Development of List Based Domain Specific Arabic Stemmers, In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools,* Cairo, Egypt.

Hu, M. and Liu, B. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD- 2004).* Seattle, Washington, USA.

Liu, B., Hu, M., and Cheng, J. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th international World Wide Web conference (WWW-2005).* Chiba, Japan

Liu, J. Wu, G. and Yao, J. 2006. Opinion Searching in Multi-Product Reviews. In *Proceedings of the Sixth IEEE International Conference on Computer and Information Technology*, 25. Seoul, Korea: IEEE Computer Society.

Miller, G., Beckwith, R., Fellbaum, C. , Gross, D. and Miller, K..1990. Introduction to WordNet: An On-line Lexical Database. *Journal of Lexicography* 3:.235-244.

Popescu A. M. and Etzioni, O. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT-EMNLP,*339–346. Vancouver, B.C., Canada.

Shi, B. and Chang, K. 2006. Mining Chinese Reviews. In *Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06),*585-589. Hong Kong, China: IEEE Computer Society.

VERCON, 2006. http://www.vercon.sci.eg.

Yi, ,J. Nasukawa, T. Bunescu, R. and Niblack, W. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of The Third IEEE International Conference on Data Mining*. Melbourne, Florida, USA.